# 語の連接の多様性に基づく日本語マニュアルからの重要語抽出

和氣 真　松崎 知美　森 辰則　中川 裕志
横浜国立大学 工学部 電子情報工学科

## 1 はじめに

最近,WINDOWSのHELP機能などで見られるような,語句をクリックする事によって現在表示されているテキストからテキストの別の部分へ表示を切り替えることが出来るようなツールが増えてきた. このようなツールの利点は,ユーザが必要な知識−例えば理解できない語句の定義−といったような情報をマウスの簡単な操作によって容易に得る事が出来る点があげられる. この利点をマニュアルに生かすと,そのマニュアル自体が非常にユーザに理解し易い便利なものになる. しかしこのようなツールを用いてマニュアルを作る際には,上で述べたようなユーザが情報を必要とする語句を予めマニュアルライターが選ばなくてはならない. しかしそのような語句の情報は客観的に抽出することが難しく,またマニュアル自体の長さが長いものになるとその抽出作業は容易なものではなくなってくる. 本稿ではそのような,ユーザが情報を必要とするであろう語句を対象マニュアルの重要語であると考え,従来の重要語抽出の文献検索システムを目的としたものとは異なる重要語の自動抽出法について述べる.

## 2 名詞の連接情報を用いた重要度計算法

### 2.1 名詞の連接情報

本稿では日本語のマニュアル文からの重要語抽出を対象とする. 日本語マニュアル文には操作法などの説明において重要な概念を示す重要語が含まれていて,そのような重要語には基本語とそれらを組み合わせた複合語ないしは名詞句があり,重要な基本語は多くの名詞と連接して多様な名詞句を作る. つまり連接する名詞の種類の数が多い名詞が重要な基本語であり,これを含む名詞句は重要語である確率が高い. 本稿では上記の考え方に基づき,名詞の連接情報を用いて重要語の候補となる名詞句の重要度を計算する手法を説明する.

### 2.2 重要度計算法

#### 2.2.1 重要語の候補語の抽出

一般に,名詞句を構成する語の品詞についての制約は,一番最後の語が名詞であることが多いと言うこと以外殆んど無い. しかし本手法ではマニュアル文における重要語を抽出するという目的から,一般にいう名詞句とは異なる名詞句を考え,それを重要語の候補となる名詞句とする.

マニュアル文における重要語の候補となる語は概ね以下のように分類できる.

- 製品もしくは製品の一部の名称
- 製品の機能
- 製品を使用する際に用いるものの名称
- 製品を使用する際に行なう動作

このような語句が重要語の候補となるため,本手法では重要語の候補となる語に以下の制約を定義する.

- 名詞を修飾する形容詞は名詞句に含まない
- 名詞を修飾する節は名詞句に含まない
- 「の」以外の助詞は含まない

以上の制約を重要語の候補となる名詞句に課すと,その構成要素は名詞及び複合名詞と助詞「の」で繋がった名詞列に限定される.

本システムはマニュアル文の形態素解析に「日本語形態素解析システムJUMAN version 1.0[JUM93]」を使用する.

JUMANの辞書において名詞に分類されている品詞は以下の6つである.

- 普通名詞
- サ変名詞
- 固有名詞
- 地名
- 人名
- 数詞

しかし実際には以上の品詞の他にも名詞となり得る未定義語がある. この未定義語とはJUMANの辞書に登録されていない語にJUMANから便宜上与えられる品詞のことであり,あらゆる品詞になる可能性がある. しかし文章をマニュアルに限定すると,未定義語になり得る語の品詞の殆んどは「動詞」「名詞」「記号」の3つである. 記号は本システム内や人間の手による処理で削除するのでここでは考えない. 動詞に関しては,マニュアル文でのサ変以外の動詞は極めて単純なものが多く辞書に定義されていない動詞が出現することはまずない. よってマニュアル文において未定義語が記号や動詞であることはなく,マニュアル文の形態素解析結果における未定義語は全て名詞であると考えて良い.

以上の考察より,重要語の候補となる名詞句は名詞に分類される語に加えて,未定義語,そして助詞「の」が連続している部分をJUMANの形態素解析の結果から抽出するものとする.

#### 2.2.2 語の連接情報の抽出

先に述べたように本稿の重要度計算は, 連接する名詞の種類の数が多い名詞が重要な基本語であり, これを含む名詞句は重要語である確率が高いという考え方に基づいて, 重要語候補の名詞句の重要度計算を行なう. ここでは重要度計算に必要な名詞の連接情報である前方連接数・後方連接数について説明する.
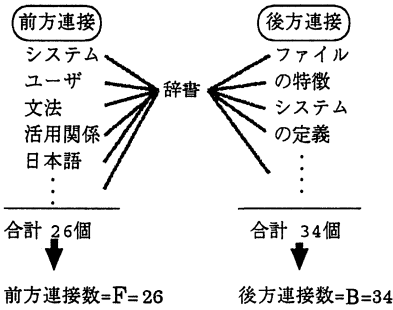
図 1: 連接数のカウント

図1は「日本語形態素解析システム JUMAN version 1.0」のマニュアル[JUM93]における名詞「辞書」の例である. このように「辞書」は, 前方に名詞を連接して「システム辞書」「ユーザ辞書」「文法辞書」「活用関係辞書」「日本語辞書」などの複合語を造り出している. この連接する名詞の種類の数をカウントして名詞「辞書」の前方連接数として定義する. 後方連接数も前方連接数と同様にカウントする.

#### 2.2.3 重要度計算式

本方式では, 複合名詞を作る力の強い名詞が重要な基本語であり, その重要な基本語を含む名詞句が重要語であるという考え方に基づいている. そこで以下のような重要度計算法を提案する.

重要語候補の名詞句を構成する名詞を前から順番に $N_1, N_2, N_3, \cdots, N_n$ とし, 名詞 $N_i$ が持つ前方連接数および後方連接数をそれぞれ $F_i, Bi$ とする. そして名詞 $N_i$ の固有重要度を $J_i$ としてそれを求める式を以下のように定義する.

$$\text{固有重要度} : J_i = \{(Fi+1) \times (Bi+1)\}^{\frac{1}{2}} \quad (1)$$

この計算によって重要語を構成している名詞 $N_1, N_2, N_3, \cdots, N_n$ はそれぞれ固有重要度 $J_1, J_2, J_3, \cdots, J_n$ を持っていることになる. そして重要度候補名詞句の構成名詞それぞれが持っている固有重要度の相乗平均をもって, その名詞句の重要度とする.

よって名詞句の重要度 $J$ は以下の式で求める.

$$\text{重要度計算式} : J = \{\prod_{i=1}^{n}\{(Fi+1) \times (Bi+1)\}\}^{\frac{1}{2n}} \quad (2)$$

### 2.3 頻度分布と重要度の関係

実験に用いたのは, 日本語形態素解析システム JUMAN の使用説明書 [JUM93] である.

相乗平均による重要度がどの程度有効であるかを調べるために, 従来よく行なわれてきた出現回数による抽出との比較を行なった. すべての名詞句のマニュアル中における出現回数を計算し, 個々の名詞句について, 出現回数を横軸に, 算出した重要度を縦軸にプロットした. その結果が図2である.
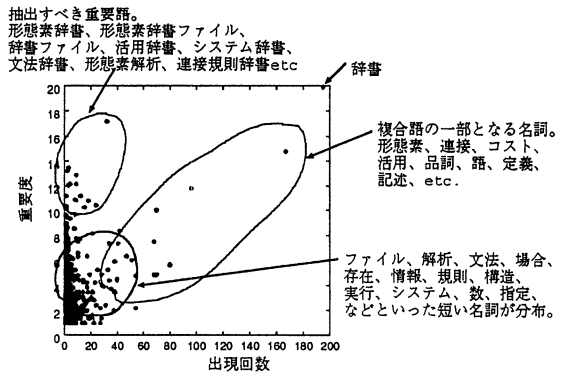
図 2: 出現回数と重要度の分布

出現回数が多い名詞というのは, どうしても, 複合語の一部となることが多いような短い名詞が多くなる. 複雑な概念が盛り込まれた複合名詞はこれらの短い名詞に比べると, 出現回数は少なくなってしまう. プロットされた具体的な名詞を見ていこう. 重要度, 出現回数ともに最大の名詞句は, 「辞書」(グラフ中一番右上のプロット)であった. グラフの原点からこの「辞書」のプロットへ向かって, 右上がりに並ぶ, 出現回数50回以上の点(図中2の領域)は, 「形態素」, 「連接」, 「コスト」, 「活用」, 「品詞」, 「語」, 「定義」, 「記述」と, 全て複合語の一部となるような短い名詞であった. このような名詞は, 出現回数が多いからといって, このまま索引語とするには適さないであろう. これに対し, 出現回数は50回以下でも, 重要度10以上, つまりグラフの左上にプロットされた名詞句(図中1の領域)は, 「形態素辞書」, 「形態素辞書ファイル」, 「辞書ファイル」, 「活用辞書」, 「システム辞書」, 「文法辞書」, 「形態素解析」, 「連接規則辞書」などである. 実際にこのシステムを利用している人を被験者としてマニュアル中から重要語を選んでもらったところ, これらの語が選ばれていた. このことからもこの計算法による重要度は意味のあるものであるといえるだろう. 図中3の領域はどうだろうか. この辺りには, 「ファイル」, 「解析」, 「文法」, 「場合」, 「存在」, 「情報」, 「規則」, 「構造」, 「実行」, 「システ

ム」，「数」，「指定」といった短い名詞が分布している．これらも，索引語にはむかない．以上まとめるとこの重要度計算は，複雑難解なマニュアルから，索引語とするのに望ましい難解な複合語，すなわち，そのマニュアル独特に頻出する概念からなる複合語を抽出する際に効果を期待できる．

## 3 重要語の絞り込み法

本研究では,SGML化した日本語マニュアル文における詳細説明のためのハイパーリンクを張るべき語句を自動抽出することを目的としている．リンクを張ると言う事はその重要語はそのマニュアル中に最低でも2つは存在しないといけない．よって対象マニュアル中に1回しか出現しない語は重要語の抽出対象から外す．これは従来の重要語抽出法の中の単語の頻度依存の抽出法においても一般的に「単語の頻度が少ないものは重要語の対象にはならない」と考えられており，この考え方は本稿の「対象マニュアル中に1回しか出現しない語は，重要語候補名詞句から外す」にも一致する．以上の考察から抽出する重要語に関して以下の制約を課する．

**重要語抽出に関する制約** 重要語は対象マニュアル中において2回以上出現している語のみを選択する

この制約に基づいて重要語候補のフィルタリングを行なう．

以上の処理を行なった後，重要語候補の名詞句を重要度の高い順にソートする．するとそのソート結果には以下のような分布の傾向が現れる．

> 固有重要度の高い名詞，すなわちそのマニュアルにおいて重要な基本語を含む名詞句は，その基本語を中心として周辺に集まる

本稿の重要度計算法では名詞句を構成する名詞の固有重要度の相乗平均を重要度としているため，固有重要度の高い名詞を含む名詞句は同様に重要度の高いところへと分布し，以上のような重要語候補名詞句の分布の傾向が現れる．

他にもマニュアル文における重要語には以下のような傾向が見られる．

- 製品もしくは製品の一部の名称
- 製品の機能
- 製品を使用する際に用いるものの名称
- 製品を使用する際に行なう動作

このような傾向を持つため重要語は複数の名詞から成る名詞句である確率が非常に高いということが言える．

上に述べた重要語に関する2つの傾向

- 重要語は重要度の高いところに，かつ1つの箇所に偏り易い
- 重要語は複数の名詞から成る名詞句である確率が高い

この2つの傾向を組み合わせて，以下の重要語抽出のための基本概念を提案する．

- 重要語は複数の名詞から成る名詞句の割合が大きい箇所の周辺にある確率が高い

この概念に基づいて以下の方法で重要語を絞り込む．

**手順1** 重要語の候補となる名詞句を重要度の順にソートする

**手順2** 手順1の結果のリストに対し，一定幅の窓を動かす

**手順3** 窓の中における複数の名詞から成る名詞句が一定割合以上あれば重要語として選ぶ[1]

図3はSAX[SAX93]の重要度計算結果の窓内における正解重要語の割合と複数の名詞から成る語の割合をグラフにしたものである．この図からも，窓内の正解重要語の割合と複数の名詞から成る語の割合には相関関係があることは明らかであり，上記の重要語の絞り込み法は有効なものであることがわかる．
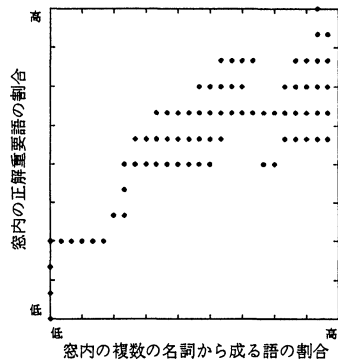
図3: 窓内における正解重要語の割合と複数の名詞から成る語の割合

## 4 抽出した重要語の評価

情報検索の評価には一般に再現率・適合率という指標を用いる．再現率は出力の洩れの少なさを，適合率は出力のノイズの少なさを示す指標となる．つまり本稿の重要語抽出で言い換えると，再現率は正解の重要語のうち何%を抽出した重要語が占めているかを示していて，その率が高いほど正解重要語の抽出洩れが少ないことになる．また適合率は抽出した重要語のう

[1]但し，重要度が上位の語，数個は無条件で抽出しておく

ち何％が正解の重要語であるかを示していて,その率が高いほど抽出した重要語の中に誤りの重要語(ノイズ)が少ないことになる.

再現率・適合率を計算するには図4に示したような正解重要語,一致重要語,抽出重要語をまず求める必要がある.
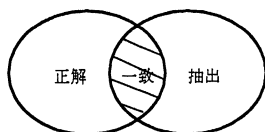
図 4: 正解・一致・抽出 重要語

図4の3種類の重要語の数をそれぞれ求め以下の計算式を用いて,適合率・再現率を計算する.

$$\text{再現率} = \frac{\text{一致重要語の数}}{\text{正解重要語の数}} \quad (3)$$

$$\text{適合率} = \frac{\text{一致重要語の数}}{\text{抽出重要語の数}} \quad (4)$$

### 4.1 人間の主観評価実験

まず正解重要語を抽出してもらう被験者には以下のような基準を設けた.

- 対象マニュアルで説明している製品を実際に使用している
- ある程度製品を理解している

この基準に当てはまる被験者数人を各マニュアル毎に募り,正解重要語抽出実験をしてもらった.

正解重要語抽出に際して,被験者には予め以下の注意事項を読んでもらった上で正解重要語を抽出してもらった.

- 単名詞もしくは複合語を1単位とし,間に助詞「の」を挟まない語句を切り離して重要語として抽出しない
- 重要語は上記の1単位,もしくは助詞「の」を挟んだ複数単位のものを重要語とする
- 重要語の基準は「その重要語に対して詳細を説明するリンクを張って欲しい語」とする

### 4.2 実験結果

本稿では以下の4つのマニュアル[2]について重要度計算および重要語選出を行ない,同様に人間の主観評価からの正解重要語を抽出した.本手法による重要語抽出は,重要語の絞り込みの窓の大きさと重要語選出条件である窓内の複数の名詞から成る名詞句の割合(表1では含有率としている)の2つの値を色々と推移させ,再現率・適合率の最適な値を求めた.それを表にしたものを表1に示す[3].

| 対象マニュアル | 窓の幅<br>含有率 | 再現率<br>適合率 |
|---|---|---|
| 日本語形態素解析システム | 12<br>4割 | 59.7<br>28.2 |
| 構文解析システム | 12<br>7割 | 60.5<br>33.2 |
| 家庭用ゲーム機 | 13<br>6割 | 81.2<br>25.1 |
| ビデオデッキ | 10<br>6割 | 61.9<br>37.1 |

表 1: 再現率・適合率

重要語(キーワード)抽出の分野において,抽出結果の再現率および適合率の最適値は共に30％前後が一般的であることから考えると,表1に示した評価結果は本稿の重要語抽出法が有効なものであることを証明している.

## 5 おわりに

本研究では,SGML化した日本語マニュアル文において詳細説明のためのハイパーリンクを張るべき語句が重要語であると考え,それを自動抽出するための手法について説明した.また実際のマニュアルを用いて重要語抽出実験を行ない,かなり有効な重要語を抽出できることが分かった.

今後の課題は重要語の絞り込みを完全に自動化することである。

## 参考文献


[JUM93] 松本裕治ら.日本語形態素解析システムＪＵＭＡＮ使用説明書version1.0. 1993.

[SAX93] 松本裕治ら.構文解析システムＳＡＸ使用説明書version2.0. 1993.

[Son] Sony Computer Entertainment Inc. PlayStation取扱説明書.

[三菱] 三菱電機株式会社. HV-F93(取扱説明書).


[2]使用したマニュアルは「日本語形態素解析システムJUMAN取扱説明書 version1.0」[JUM93]「構文解析システムＳＡＸ使用説明書version2.0」[SAX93]「PlayStation取扱説明書」[Son]「HV-F93(取扱説明書)」(ビデオデッキのマニュアル)[三菱]の4つである.

[3]便宜上,再現率・適合率の値は百分率で表した